School of Informatics and Computing at Indiana University
Background of the School

- The School of Informatics was established in 2000 as first of its kind in the United States.
- Computer Science was established in 1971 and became part of the school in 2005.
- Library and Information Science was established in 1951 and became part of the school in 2013.
- Now named the School of Informatics and Computing.
What Is Our School About?

The broad range of computing and information technology: science, a broad range of applications and human and societal implications.

United by a focus on information and technology, our extensive programs include:

- Computer Science
- Informatics
- Information Science
- Library Science
- Data Science (starting)
Size of School (2013-2014)

- Faculty: 97 (85 tenure track)
- Students:
  - Undergraduate: 1,191
  - Master’s: 644
  - Ph.D.: 263
- Female Undergraduates: 21% (↑68% since 2007)
- Female Graduate Students: 28% (↑4% since 2007)

Undergraduates mainly Informatics; Graduates mainly Computer Science
Data Science Cosmically
There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.

Perhaps Informatics/ILS aimed at 1.5 million jobs. Computer Science covers the 140,000 to 190,000

Big Data about an order of magnitude larger than data science

21 September 2014
15,639 jobs have “big data” phrase
What is Data Science?

• The next slide gives a definition arrived by a NIST study group fall 2013.

• The previous slide says there are several jobs but that’s not enough! Is this a field – what is it and what is its core?

– The emergence of the 4th or data driven paradigm of science illustrates significance - http://research.microsoft.com/en-us/collaboration/fourthparadigm/
  – Discovery is guided by data rather than by a model
  – The End of (traditional) science http://www.wired.com/wired/issue/16-07 is famous here

• Another example is recommender systems in Netflix, e-commerce etc.
  – Here data (user ratings of movies or products) allows an empirical prediction of what users like
  – Here we define points in spaces (of users or products), cluster them etc. – all conclusions coming from data
Data Science Definition from NIST Public Working Group

- **Data Science** is the extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and analytical hypothesis analysis.

- A **Data Scientist** is a practitioner who has sufficient knowledge of the overlapping regimes of expertise in business needs, domain knowledge, analytical skills and programming expertise to manage the end-to-end scientific method process through each stage in the big data lifecycle.

See Big Data Definitions in http://bigdatawg.nist.gov/V1_output_docs.php
Some Existing Online Data Science Activities

• Indiana University is “blended”: online and/or residential; other universities offer residential

• We may discount online when total cost ~$11,500 (in state price)

<table>
<thead>
<tr>
<th>Institution</th>
<th>Program Name</th>
<th>Total Tuition</th>
<th>Cost Per Credit Hour</th>
<th>Credit Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC- Berkeley</td>
<td>Master of Science in Information &amp; Data Science</td>
<td>$59,994</td>
<td>$2,222</td>
<td>27</td>
</tr>
<tr>
<td>Northwestern University</td>
<td>Master of Science in Predictive Analytics</td>
<td>$46,404</td>
<td>$1,289</td>
<td>36</td>
</tr>
<tr>
<td>Indiana University</td>
<td>Master of Science in Data Science**</td>
<td>$35,490</td>
<td>$1,183</td>
<td>30</td>
</tr>
<tr>
<td>DePaul University</td>
<td>Master of Science in Predictive Analytics</td>
<td>$40,820</td>
<td>$785</td>
<td>52</td>
</tr>
<tr>
<td>Lewis University</td>
<td>Master of Science in Data Science</td>
<td>$27,360</td>
<td>$760</td>
<td>36</td>
</tr>
<tr>
<td>CUNY - City College</td>
<td>Master of Science in Data Analytics</td>
<td>$26,820</td>
<td>$745</td>
<td>36</td>
</tr>
<tr>
<td>University of Maryland University College</td>
<td>Master of Science in Data Analytics</td>
<td>$24,984</td>
<td>$694</td>
<td>36</td>
</tr>
<tr>
<td>Elmhurst College</td>
<td>Master of Science in Data Science</td>
<td>$23,250</td>
<td>$775</td>
<td>30</td>
</tr>
<tr>
<td>Southern New Hampshire University</td>
<td>Master of Science in Data Analytics</td>
<td>$22,572</td>
<td>$627</td>
<td>36</td>
</tr>
</tbody>
</table>
Data Science Curriculum at Indiana University

Faculty in Data Science is “virtual department”

4 course Certificate: purely online, started January 2014

10 course Masters: online/residential, expected to start January 2015
GRADUATE

Data Science

What makes IU's Data Science different

We believe there are four significant ways in which our science program stands out from our competitors: coverage, world-leading faculty, strong industry links, excellent career services, and cost. Our courses are developed and taught by faculty who are international leaders in their fields. The School of Informatics and Computing is one of the pioneers in the next generation of computing, with areas of expertise as diverse as complex systems, high performance computing, networks and visualization, bioinformatics. We have a strong record of student success, including research opportunities. As a student in the data science program, you get full access to the resources of our school, including our research faculty, strong industry links, and opportunities to progress further in industry careers. All this for a total cost of only $4,500.

Hot Jobs

IBM estimates that over 90 percent of the data in the world has been created since 2011. Data science is an emerging field as a result of this massive data growth. According to McKinsey Global Institute report, by 2018 the U.S. will face a shortage of 140,000 to 190,000 people with analytical skills, as well as 1.5 million managers and with the know-how to use the analysis of big data to make effective decisions. As a student in the data science program, you have access to the School’s dynamic career services, giving you a head start in finding the first job or career you are right for.

Online Certificate in Data Science

Apply to the Online Data Science Program

Contact Us
Indiana University Data Science Certificate

• We currently have **75 students admitted into the Data Science Certificate program** (from 81 applications)
• **36** students admitted in Spring 2014; **17** of these have signed up for fall classes
• **39** students admitted in Fall 2014; **35** of these have signed up for fall classes and **4** are in process
• We expected many more applicants
• Two tracks for information only
  – Decision Maker (little software) ~ = McKinsey “managers and analysts”
  – Technical ~ = McKinsey “people with deep analytical skills”
• Total tuition costs for the twelve credit hours for this certificate is approximately $4,500. (Factor of three lower than out of state $14,198 and ~ in-state rate $4,603)
Data Science Masters Features

• Fully approved by University; expected to be approved by State October 2014
• Blended online and residential
• Department of Information and Library Science, Division of Informatics and Division of Computer Science in the Department of Informatics and Computer Science, School of Informatics and Computing and the Department of Statistics, College of Arts and Science, IUB
• 30 credits (10 conventional courses)
• Basic (general) Masters degree plus tracks
  – Currently only track is “Computational and Analytic Data Science”
  – Other tracks can be proposed and approved by campus, data science faculty, data science curriculum committee
• A 4-course Certificate in Data Science has been approved, and a Ph.D. Minor in Data Science will be proposed.
3 Types of Students

• Professionals wanting skills to improve job or “required” by employee to keep up with technology advances
• Traditional sources of IT Masters
• Students in non IT fields wanting to do “domain specific data science”
What do students want?

• Degree with some relevant curriculum
  – Data Science and Computer Science distinct BUT

• Real goal often “Optional Practical Training” OPT allowing graduated students visa to work for US companies
  – Must have spent at least a year in US in residential program

• Residential CS Masters (at IU) 95% foreign students

• Online program students quite varied but mostly USA professionals aiming to improve/switch job
IU and Competition

• With Computer Science, Informatics, ILS, Statistics, IU has particularly broad unrivalled technology base
  – Other universities have more domain data science than IU

• Existing Masters in US in table. Many more

<table>
<thead>
<tr>
<th>School</th>
<th>Program</th>
<th>Campus</th>
<th>Online</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columbia University</td>
<td>Data Science</td>
<td>Yes</td>
<td>No</td>
<td>MS 30 cr</td>
</tr>
<tr>
<td>Illinois Institute of Technology</td>
<td>Data Science</td>
<td>Yes</td>
<td>No</td>
<td>MS 33 cr</td>
</tr>
<tr>
<td>New York University</td>
<td>Data Science</td>
<td>Yes</td>
<td>No</td>
<td>MS 36 cr</td>
</tr>
<tr>
<td>University of California Berkeley School of Information</td>
<td>Master of Information and Data Science</td>
<td>Yes</td>
<td>Yes</td>
<td>M.I.D.S</td>
</tr>
<tr>
<td>University of Southern California</td>
<td>Computer Science with Data Science</td>
<td>Yes</td>
<td>No</td>
<td>MS 27 cr</td>
</tr>
</tbody>
</table>
Basic Masters Course Requirements

• One course from two of three technology areas
  – I. Data analysis and statistics
  – II. Data lifecycle (includes “handling of research data”)
  – III. Data management and infrastructure
• One course from (big data) application course cluster
• Other courses chosen from list maintained by Data Science Program curriculum committee (or outside this with permission of advisor/ Curriculum Committee)
• Capstone project optional
• All students assigned an advisor who approves course choice.
• Due to variation in preparation will label courses
  – Decision Maker
  – Technical
• Corresponding to two categories in McKinsey report – note Decision Maker had an order of magnitude more job openings expected
Computational and Analytic Data Science track

• For this track, data science courses have been reorganized into categories reflecting the topics important for students wanting to prepare for computational and analytic data science careers for which a strong computer science background is necessary. Consequently, students in this track must complete additional requirements,

• 1) A student has to take at least 3 courses (9 credits) from **Category 1 Core Courses**. Among them, B503 Analysis of Algorithms is required and the student should take at least 2 courses from the following 3:
  – B561 Advanced Database Concepts,
  – [STAT] S520 Introduction to Statistics OR (New Course) Probabilistic Reasoning
  – B555 Machine Learning OR I590 Applied Machine Learning

• 2) A student must take at least 2 courses from **Category 2 Data Systems**, AND, at least 2 courses from **Category 3 Data Analysis**. Courses taken in Category 1 can be double counted if they are also listed in Category 2 or Category 3.

• 3) A student must take at least 3 courses from Category 2 Data Systems, OR, at least 3 courses from Category 3 Data Analysis. Again, courses taken in Category 1 can be double counted if they are also listed in Category 2 or Category 3. One of these courses must be an application domain course.
Comparing Google Course Builder (GCB) and Microsoft Office Mix
<table>
<thead>
<tr>
<th>Course Content</th>
<th>Units</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Section 1 - Introduction</td>
<td>1, 2</td>
<td>2h 37min</td>
</tr>
<tr>
<td>+ Section 2 - Overview of Data Science: What is Big Data, Data Analytics and X-Informatics?</td>
<td>3, 4, 5</td>
<td>2h 32min</td>
</tr>
<tr>
<td>+ Section 3 - Technology Training - Python &amp; FutureGrid</td>
<td>6, 7</td>
<td>1h 53min</td>
</tr>
<tr>
<td>+ Section 4 - X= Physics Case Study</td>
<td>8, 9, 10, 11</td>
<td>3h 7min</td>
</tr>
<tr>
<td>+ Section 5 - Big Data Use Cases Survey</td>
<td>12, 13, 14</td>
<td>5h 18min</td>
</tr>
<tr>
<td>+ Section 6 - Technology Training - Plotviz</td>
<td>15</td>
<td>1h</td>
</tr>
<tr>
<td>+ Section 7 - X= e-Commerce and Lifestyle Case Study</td>
<td>16, 17, 18</td>
<td>2h 12min</td>
</tr>
<tr>
<td>+ Section 8 - Technology Training - kNN &amp; Clustering</td>
<td>19, 20</td>
<td>1h 23min</td>
</tr>
<tr>
<td>+ Section 9 - Cloud Computing Technology for Big Data Applications &amp; Analytics</td>
<td>21, 22, 23, 24</td>
<td>3h 6min</td>
</tr>
<tr>
<td>+ Section 10 - X-Informatics with X = Web Search and Text Mining and their technologies</td>
<td>25, 26</td>
<td>1h 39min</td>
</tr>
<tr>
<td>+ Section 11 - Technology for Big Data Applications &amp; Analytics</td>
<td>27, 28, 29, 30</td>
<td>1h 58min</td>
</tr>
<tr>
<td>+ Section 12 - X= Health Informatics Case Study</td>
<td>31</td>
<td>1h 2min</td>
</tr>
<tr>
<td>+ Section 13 - X= Sensors Case Study</td>
<td>32</td>
<td>32min</td>
</tr>
<tr>
<td>+ Section 14 - X= Radar Case Study</td>
<td>33</td>
<td>21min</td>
</tr>
</tbody>
</table>
This course INFO 590 credit course investigates the use of cloud running data analytics collaboratively for processing Big Data to solve problems in Big Data Applications and Analytics. Case studies such as Netflix recommender systems, Genomic data, and more will be discussed.

1. COURSE SUMMARY

INF 590 Graduate class is ONLY for Data Science graduate track non-residents that contains further enhanced content, Java and Python, and conventional grading using OnCourse.

Data Science Graduate class for remote students: further enhanced content, Java and Python, INF course credit with normal grading.

2. MORE INFO

Homework for your class will be posted in OnCourse. Grading will also be done conventionally using INF OnCourse. Course Material: syllabus, course files, slides, or all-in-one zip file. Use the Google Community forum for course discussions.

Once you add yourself to the community, we will approve your request after checking your enrollment.

Please enroll only using a GMAIL account

Instructor

Professor Geoffrey Fox received a PhD in Theoretical Physics from Cambridge University and is now Professor of Informatics and Computing as well as Physics at Indiana University, where he is director of the Digital Science Center and Associate Dean for Research and Graduate Studies at the School of Informatics and Computing. He previously held positions at Caltech, Syracuse University and Florida State University.

He has published around 1,000 papers in Physics and Computer Science, supervised the PhD candidacies of 65 students, and received an h-Index of 67 along with over 23000 citations.

Professor Fox currently works in applying Computer Science to Bioinformatics, Sensor Clouds, Earthquake and Ice-sheet Forecast, and Public Health. He is a founder of CTO of

3. OTHER INFO

Course Syllabus (PDF)
Create video in PowerPoint with laptop web cam

Exported to Microsoft Video Streaming Site
Course description
This course studies software used in many commercial activities to study Big Data. The backdrop for course is the ~120 software subsystems illustrated here. We will describe the software architecture represented by this collection which we term HPC-ABDS (High Performance Computing enhanced Apache Big Data Stack).

A paper discussing this can be found here and presentations here and here. Copies of this material may be found in resources. You can download the syllabus (PDF).

Lessons

Unit 1: Introduction
1. Part A
2. Part B
3. Part C

Unit 2: Data Access Patterns and Introduction to using HPC-ABDS
1. Part A
2. Part B
3. Part C
4. Part D

Unit 3: Big Data Applications Structure

Office Mix Site

Lectures

Made as ~15 minute lessons linked here

Metadata on Microsoft Site
The lessons on my Microsoft Site


By Geoffrey Fox

Created 1 week ago | Duration 0:15:09  1

Unit 14E(1590HPCABDSMooocFall2014-014Ev2): Big Data Applications and Generalizing their Structure Part E of 1590: Two facets of the Big Data Ogres. Data source and Style plus Data analytics kernels. Summary of this entire unit: parts A B C D E


By Geoffrey Fox

Created 1 week ago | Duration 0:22:44  1

Google Community Group
Potpourri of Online Technologies

- **Canvas (Indiana University Default):** Best for interface with IU grading and records
- **Google Course Builder:** Best for management and integration of components
- **Ad hoc web pages:** alternative easy to build integration
- **Mix:** Best faculty preparation interface
- **Adobe Presenter/Camtasia:** More powerful video preparation that support subtitles but not clearly needed
- **Google Community:** Good social interaction support
- **YouTube:** Best user interface for videos
- **Hangout:** Best for instructor-students online interactions (one instructor to 9 students with live feed). Hangout on air mixes live and streaming (30 second delay from archived YouTube) and more participants
Details of Masters Degree
Computational and Analytic Data Science track

• **Category 1: Core Courses**
  - CSCI B503 Analysis of Algorithms
  - CSCI B555 Machine Learning OR INFO I590 Applied Machine Learning
  - CSCI B561 Advanced Database Concepts
  - STAT S520 Introduction to Statistics OR (New Course) Probabilistic Reasoning

• **Category 2: Data Systems**
  - CSCI B649 Cloud Computing CSCI B649 Advanced Topics in Privacy
  - CSCI P538 Computer Networks
  - INFO I533 Systems & Protocol Security & Information Assurance
  - ILS Z534: Information Retrieval: Theory and Practice
Computational and Analytic Data Science track

- **Category 3: Data Analysis**
  - CSCI B565 Data Mining
  - CSCI B555 Machine Learning
  - INFO I590 Applied Machine Learning
  - INFO I590 Complex Networks and Their Applications
  - STAT S520 Introduction to Statistics
  - (New Course) Probabilistic Reasoning
  - (New Course CSCI) Algorithms for Big Data

- **Category 4: Elective Courses**
  - CSCI B551 Elements of Artificial Intelligence
  - CSCI B553 Probabilistic Approaches to Artificial Intelligence
  - CSCI B659 Information Theory and Inference
  - CSCI B661 Database Theory and Systems Design
  - INFO I519 Introduction to Bioinformatics
  - INFO I520 Security For Networked Systems
  - INFO I529 Machine Learning in Bioinformatics
  - INFO I590 Relational Probabilistic Models
  - ILS Z637 - Information Visualization
  - Every course in 500/600 SOIC related to data that is not in the list
  - All courses from STAT that are 600 and above
Admissions

• Decided by Data Science Program Curriculum Committee

• Need some computer programming experience (either through coursework or experience), and a mathematical background and knowledge of statistics will be useful

• Tracks can impose stronger requirements

• 3.0 Undergraduate GPA

• A 500 word personal statement

• GRE scores are required for all applicants.

• 3 letters of recommendation
Four Areas I and II

• **I. Data analysis and statistics:** gives students skills to develop and extend algorithms, statistical approaches, and visualization techniques for their explorations of large scale data. Topics include data mining, information retrieval, statistics, machine learning, and data visualization and will be examined from the perspective of “big data,” using examples from the application focus areas described in Section IV.

• **II. Data lifecycle:** gives students an understanding of the data lifecycle, from digital birth to long-term preservation. Topics include data curation, data stewardship, issues related to retention and reproducibility, the role of the library and data archives in digital data preservation and scholarly communication and publication, and the organizational, policy, and social impacts of big data.
Four Areas III and IV

• **III. Data management and infrastructure:** gives students skills to manage and support big data projects. Data have to be described, discovered, and actionable. In data science, issues of scale come to the fore, raising challenges of storage and large-scale computation. Topics in data management include semantics, metadata, cyberinfrastructure and cloud computing, databases and document stores, and security and privacy and are relevant to both data science and “big data” data science.

• **IV. Big data application domains:** gives students experience with data analysis and decision making and is designed to equip them with the ability to derive insights from vast quantities and varieties of data. The teaching of data science, particularly its analytic aspects, is most effective when an application area is used as a focus of study. The degree will allow students to specialize in one or more application areas which include, but are not limited to Business analytics, Science informatics, Web science, Social data informatics, Health and Biomedical informatics.
I. Data Analysis and Statistics

- CSCI B503 Analysis of Algorithms
- CSCI B553 Probabilistic Approaches to Artificial Intelligence
- CSCI B652: Computer Models of Symbolic Learning
- CSCI B659 Information Theory and Inference
- CSCI B551: Elements of Artificial Intelligence
- CSCI B555: Machine Learning
- CSCI B565: Data Mining
- INFO I573: Programming for Science Informatics
- INFO I590 Visual Analytics
- INFO I590 Relational Probabilistic Models
- INFO I590 Applied Machine Learning
- ILS Z534: Information Retrieval: Theory and Practice
- ILS Z604: Topics in Library and Information Science: Big Data Analysis for Web and Text
- ILS Z637: Information Visualization
- STAT S520 Intro to Statistics
- STAT S670: Exploratory Data Analysis
- STAT S675: Statistical Learning & High-Dimensional Data Analysis
- (New Course CSCI) Algorithms for Big Data
- (New Course CSCI) Probabilistic Reasoning
II. Data Lifecycle

- INFO I590: Data Provenance
- INFO I590 Complex Systems
- ILS Z604 Scholarly Communication
- ILS Z636: Semantic Web
- ILS Z652: Digital Libraries
- ILS Z604: Data Curation
- (New Course INFO): Social and Organizational Informatics of Big Data
- (New Course ILS): Project Management for Data Science
- (New Course ILS): Big Data Policy
III. Data Management and Infrastructure

- CSCI B534: Distributed Systems
- CSCI B552: Knowledge-Based Artificial Intelligence
- CSCI B561: Advanced Database Concepts
- CSCI B649: Cloud Computing (offered online)
- CSCI B649 Advanced Topics in Privacy
- CSCI B649: Topics in Systems: Cloud Computing for Data Intensive Sciences
- CSCI B661: Database Theory and System Design
- CSCI B662 Database Systems & Internal Design
- CSCI B669: Scientific Data Management and Preservation
- CSCI P536: Operating Systems
- CSCI P538 Computer Networks
- INFO I520 Security For Networked Systems
- INFO I525: Organizational Informatics and Economics of Security
- INFO I590 Complex Networks and their Applications
- INFO I590: Topics in Informatics: Data Management for Big Data
- INFO I590: Topics in Informatics: Big Data Open Source Software and Projects
- ILS S511: Database
- Every course in 500/600 SOIC related to data that is not in the list
IV. Application areas

• CSCI B656: Web mining
• CSCI B679: Topics in Scientific Computing: High Performance Computing
• INFO I519 Introduction to Bioinformatics
• INFO I529 Machine Learning in Bioinformatics
• INFO I533 Systems & Protocol Security & Information Assurance
• INFO I590: Topics in Informatics: Big Data Applications and Analytics
• INFO I590: Topics in Informatics: Big Data in Drug Discovery, Health and Translational Medicine
• ILS Z605: Internship in Data Science
• Kelley School of Business: business analytics course(s)
• Other courses from Indiana University e.g. Physics Data Analysis
Technical Track of General DS Masters

• Year 1 Semester 1:
  – INFO 590: Topics in Informatics: Big Data Applications and Analytics
  – ILS Z604: Big Data Analytics for Web and Text
  – STAT S520: Intro to Statistics

• Year 1: Semester 2:
  – CSCI B661: Database Theory and System Design
  – ILS Z637: Information Visualization
  – STAT S670: Exploratory Data Analysis

• Year 1: Summer:

• Year 2: Semester 3:
  – CSCI B555: Machine Learning
  – STAT S670: Exploratory Data Analysis
  – CSCI B649: Cloud Computing
Computational and Analytic Data Science track

• Year 1 Semester 1:
  – B503 Analysis of Algorithms
  – B561 Advanced Database Concepts
  – S520 Introduction to Statistics

• Year 1: Semester 2:
  – B649 Cloud Computing
  – Z534: Information Retrieval: Theory and Practice
  – B555 Machine Learning

• Year 1: Summer:
  – ILS 605: Internship in Data Science

• Year 2: Semester 3:
  – B565 Data Mining
  – I520 Security For Networked Systems
  – Z637 - Information Visualization
An Information-oriented Track

- **Year 1 Semester 1:**
  - INFO 590: Topics in Informatics: Big Data Applications and Analytics
  - ILS Z604 Big Data Analytics for Web and Text.
  - STAT S520 Intro to Statistics

- **Year 1: Semester 2:**
  - CSCI B661 Database Theory and System Design
  - ILS Z637: Information Visualization
  - ILS Z653: Semantic Web

- **Year 1: Summer:**
  - ILS 605: Internship in Data Science

- **Year 2: Semester 3:**
  - ILS Z604 Data Curation
  - ILS Z604 Scholarly Communication
  - INFO I590: Data Provenance
MOOC’s

The MOOC version of Big Data Applications and Analytics has ~2000 students enrolled.

Coursera Offerings are much larger enrollment
Background

• MOOC’s are a “disruptive force” in the educational environment
  – Coursera, Udacity, Khan Academy and many others
• MOOC’s have courses and technologies
• Google Course Builder and OpenEdX are open source MOOC technologies
• Blackboard and others are learning management systems with (some) MOOC support
MOOC Style Implementations

• Courses from commercial sources, universities and partnerships

• Courses with 100,000 students (free)

• Georgia Tech a leader in rigorous academic curriculum – MOOC style Masters in Computer Science (pay tuition, get regular GT degree)

• Indiana University a much more modest Data Science certificate with 4 MOOC courses Spring 2014

• Interesting way to package tutorial material for computers and software e.g.
  – FutureGrid has had 24 EOT projects over last year (semester courses to workshops)
  – Support by MOOC modules on how to use FutureGrid
Example Google Course Builder MOOC

4 levels
Course Section (12)
Units(29)
Lessons(~150)

Units are ~ traditional lecture
Lessons are ~10 minute segments

http://x-informatics.appspot.com/course
Example Google Course Builder MOOC

The Physics Section expands to 4 units and 2 Homeworks

Unit 9 expands to 5 lessons

Lessons played on Youtube “talking head video + PowerPoint”

http://x-informatics.appspot.com/course
Random Variables

• The area of probability and statistics underlies the analysis of physics data and in fact historically the first large analyses of this type were done in physics
  – Now there are more examples and physics is not the largest
• The physics experiment observes proton-proton collisions at LHC and look at a “random” subset of possible results. The “theory” (which cannot be calculated) gives a probability for each final state occurs.
• The experimental results are sets of follow-up (called events) which form a random variable represented by a collection of individual measurements
  – Events – isolated independent activities are very important e.g. earthquakes, stock and bank transactions are events
• The DIKW pipeline transforms this random variable into different forms with final random variable having one component – does this event fall in bin of histogram
IP-over-P2P (IPOP)

Intro to FutureGrid

Cloud Computing Tutorial at Mini-PRAGMA 2013
Intro to FutureGrid MOOC: Overview 1: Overall Introduction

Prof. Geoffrey Fox introduces Future Grid in terms of its services, features and future directions.
MOOCs in SC community

• Activities like CI-Tutor and HPC University are community activities that have collected much re-usable education material.

• MOOC’s naturally support re-use at lesson or higher level
  – e.g. include MPI on XSEDE MOOC as part of many parallel programming classes.

• Need to develop agreed ways to use backend servers (HPC or Cloud) to support MOOC laboratories
  – Students should be able to take MOOC classes from tablet or phone.

• Parts of MOOC’s (Units or Sections) can be used as modules to enhance classes in outreach activities.
<table>
<thead>
<tr>
<th>Unit</th>
<th>Title</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>Geoffrey Fox</td>
</tr>
<tr>
<td>2</td>
<td>Biology in the Cloud</td>
<td>Michael Schatz</td>
</tr>
<tr>
<td>3</td>
<td>Infrastructure Used: FutureGrid</td>
<td>Geoffrey Fox</td>
</tr>
<tr>
<td>4</td>
<td>Virtualization on HPC</td>
<td>Thomas J. Hacker</td>
</tr>
<tr>
<td>5</td>
<td>Running MapReduce in Non-Traditional Environments</td>
<td>Abhishek Chandra</td>
</tr>
<tr>
<td>6</td>
<td>Virtual Clusters Supporting MapReduce in Cloud</td>
<td>Jonathan Klinginsmith</td>
</tr>
<tr>
<td>7</td>
<td>Hadoop and HDFS</td>
<td>Jerome Mitchell</td>
</tr>
<tr>
<td>8</td>
<td>HBase and Bigtable Storage</td>
<td>Thilina Gunarathe</td>
</tr>
<tr>
<td>9</td>
<td>Data Mining with Twister Iterative MapReduce</td>
<td>Hui Li</td>
</tr>
<tr>
<td>10</td>
<td>Twister Introduction</td>
<td>Judy Qiu</td>
</tr>
<tr>
<td>11</td>
<td>Commercial IaaS/PaaS I: AWS: The Platform for Data Science</td>
<td>Jamie Kinney</td>
</tr>
<tr>
<td>12</td>
<td>Commercial IaaS/PaaS II: Azure and Twister4Azure</td>
<td>Thilina Gunarathe</td>
</tr>
<tr>
<td>13</td>
<td>Federating HPC, Cyberinfrastructure and Clouds using CometCloud I</td>
<td>Manish Parashar</td>
</tr>
<tr>
<td>14</td>
<td>Scientific Workflows in the Cloud</td>
<td>Gideon Juve</td>
</tr>
<tr>
<td>15</td>
<td>Cloud Technology: Virtual Private Clusters: Virtual Appliances and Networks in the Cloud</td>
<td>Renato Figueiredo</td>
</tr>
<tr>
<td>16</td>
<td>Clouds and the Network</td>
<td>Martin Swany</td>
</tr>
<tr>
<td>17</td>
<td>Applications of Cloud: DOE Systems Biology Knowledgebase</td>
<td>Rick Stevens</td>
</tr>
<tr>
<td>18</td>
<td>Cloud Technology: Cloud Security: New Challenges and New Opportunities</td>
<td>XiaoFeng Wang</td>
</tr>
<tr>
<td>19</td>
<td>Magellan: Evaluating Cloud Computing for Science</td>
<td>Lavanya Ramakrishnan</td>
</tr>
<tr>
<td>20</td>
<td>Cloud Technology: GPU on Clouds</td>
<td>Andrew J. Younge</td>
</tr>
</tbody>
</table>
Structure of Google Course Builder (GCB) Course
Structure of GCB Course I

- 3 for-credit sections: Undergraduate, graduate, Online Data Science Certificate plus an older free MOOC
- A online course resource built with Google Course Builder and enhancements CGL Mooc Builder [http://moocbuilder.soic.indiana.edu/](http://moocbuilder.soic.indiana.edu/) built by us and available as open source that allow convenient assembly of the different course components. These components include
  - 5-15 minute video segments called lessons and containing curricula material (instructor desktop often containing PowerPoint slides).
  - Lessons are assembled into units totaling around 45 minutes – 2 hours and roughly equivalent to a traditional class.
  - Units linked into sections that together make up a coherent description of a major topic in course; for example “introduction” “Big Data and the Higgs Boson” and “Cloud Technology” are sections in these classes
Structure of GCB Course II

- The 3 sections share the same online site with 14 sections; 33 units and 220 lessons totaling 28.7 hours of video. The average lesson length was 7.8 minutes with 52 minute average for units and sections averaging just over 2 hours with a maximum length of 5 hours 18 minutes. Offering 1) was similar but had earlier versions of material.

- Each lesson had a video located on YouTube and an abstract (called lesson overview in figure 1 below). This interface show all lessons (13) for this unit and that each unit has its own abstract and slides available. There are also a list of follow-up resources associated with units and illustrated at bottom of figure 1. In the middle of figure 1, one sees the link to YouTube hosting of this lesson and 3 discussion links; one for each offering 2), 3) and 4). These are described later.
A typical lesson (the first in unit 13) Note links to all units across the top (29 of 33 units)
Course Home Page with Overview material

1. COURSE SUMMARY

This course INFO 400 and INFO 590 credit course investigates the use of cloud computing and big data analytics collaboratively for processing Big Data to solve problems in Big Data Applications and Analytics. Case studies such as Netflix recommendor systems, Genomic data, and more will be discussed.

2. FOR IU STUDENTS

- INFO 400 Undergraduate class for residents will have enhanced content, Python, conventional grading.
- INFO 590 Graduate class for residents: further enhanced content, Java and Python, conventional grading.
- Data Science Graduate class for remote students: further enhanced content, Java and Python, IU course credit with normal grading.

3. OTHER INFO

Homework for your class will be posted in Course. Grading will be done conventionally using IU Course. Course Material: syllabus, course files, slides, or all-in-one zip file. Use the appropriate Google community forum for course discussions. INFO-400, INFO-590, and INFO-DSC.

Please enroll only using a Gmail account.

Instructor

Professor Geoffrey Fox received a PhD in Theoretical Physics from Cambridge University and is now Professor of Informatics and Computing as well as Physics at Indiana University, where he is director of the Digital Science Center and Associate Dean for Research and Graduate Studies at the School of Informatics and Computing. He previously held positions at Caltech, Syracuse University and Florida State University.

He has published around 1,000 papers in Physics and Computer Science, supervised the PhD candidates of 65 students, and received an h-index of 67 along with over 23,000 citations. Professor Fox currently works in applying...
Note that we have a course – section – unit – lesson hierarchy (supported by Mooc Builder) with abstracts available at each level of hierarchy. The home page has overview information (shown earlier) plus a list of all sections and a syllabus shown above.
Figure shows a partial list of sections showing how one can interactively browse the hierarchy. The next level would expose an individual unit.
Homeworks

• These are online within Google Course Builder for the MOOC with peer assessment. In the 3 credit offerings, all graded material (homework and projects) is conducted traditionally through Indiana University Oncourse (superceded by Canvas).

• Oncourse was additionally used to assign which videos should be watched each week and the discussion forum topics described later (these were just “special homeworks in Oncourse). In the non-residential data science certificate class, the students were on a variable schedule (as typically working full time and many distractions; one for example had faculty position interviews) and considerable latitude was given for video and homework completion dates.
Discussion Forums

• Each offering had a separate set of electronic discussion forums which were used for class announcements (replicating Oncourse) and for assigned discussions. Figure 5 illustrates an assigned discussion on the implications of the success of e-commerce for the future of “real malls”. The students were given “participation credit” for posting here and these were very well received. Our next offering will make greater use of these forums. Based on student feedback we will encourage even greater participation through students both posting and commenting. Note I personally do not like specialized (walled garden) forums and the class forums were set up using standard Google Community Groups with a familiar elegant interface. These community groups also link well to Google Hangouts described later.

• As well as interesting topics, all class announcements were made in the “Instructor” forum repeating information posted at Oncourse. Of course no sensitive material such as returned homework was posted on this site.
Hangouts

• For the purely online offering, we supplemented the asynchronous material described above with real-time interactive Google Hangout video sessions illustrated in figure 6. Given varied time zones and weekday demands on students, these were held at 1pm Eastern on Sundays. Google Hangouts are conveniently scheduled from community page and offer interactive video and chat capabilities that were well received. Other technologies such as Skype are also possible. Hangouts are restricted to 10-15 people which was sufficient for this course. Not all of 12 students attended a given class. The Hangouts focused on general data science issues and the mechanics of the class.
Figure 5: The community group for one of classes and one forum (“No more malls”) described in text.
Figure 6: Community Events for Online Data Science Certificate
In class Sessions

• The residential sections had regular in class sessions; one 90 minute session per class each week. This was originally two sessions but reduced to one partly because online videos turned these into “flipped classes” with less need for in class time and partly to accommodate more students (77 total graduate and undergraduate). These classes were devoted to discussions of course material, homework and largely the discussion forum topics. This part of course was not greatly liked by the students – especially the undergraduate section which voted in favor of a model with only the online components (including the discussion forums which they recommended expanding). In particular the 9.30am start time was viewed as too early and intrinsically unattractive.